Multivariate Statistical Analysis, Exercise 1, Fall 2011, Prof. S.C. Jeng

October 7, 2011 TA: H.C. Cheng

(Geometrical Interpretation of Statistics)

From an experiment we obtain a size N random sample $\mathbf{x}_1, \ldots \mathbf{x}_N$ from some unknown population $f(\mathbf{x})$, where each random sample $\mathbf{x} = \begin{bmatrix} x_1 & \cdots & x_p \end{bmatrix}$ contains p variables. The random sample can be formulated as the following matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \cdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{bmatrix},$$

where in this experiment the number of samples N equals 10000 and the number of variables of each sample data $p = 2$. Please answer the following questions.

(a) Plot the locations of the random sample on the $\mathbb{R}^2$ space. Generate the histogram of the sample data and see its distributions.

- Solution: See Fig. 1, 2, 3, 4, 5.

(b) Calculate the sample mean $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$, sample variance $s_{ii} = \frac{1}{N-1}\sum_{j=1}^{N}(x_{ji} - \bar{x}_i),\ i = 1, 2$, and the sample correlation coefficient $r_{x_1x_2} = \frac{\sum_{j=1}^{N}(x_{j1}-\bar{x}_1)(x_{j2}-\bar{x}_2)}{(N-1)}$. Are the variables of the population statistically independent? (Assume that the measurement error and the noise contaminating the random sample is sufficiently small such that we can judge the statistical independence based on the statistics of random sample.)

- Solution: The statistics and the true parameters of the population are listed in the table below.

Statistics	Realization	True Parameters	Error
Sample Mean x_1	-0.7896	-0.8	-0.0130
Sample Mean x_2	0.4118	0.4	0.0295
Sample Variance s_{11}	0.8012	0.8	0.0015
Sample Variance s_{22}	1.0858	1.1	-0.0129
Sample Correlation Coefficient $r_{x_1x_2}$	0.8025	0.8	0.0031

Since the sample correlation coefficient is far away from zero, the two variables of the random sample is not statistically independent under the sufficient small measurement error.

The probability density function (pdf) of the population is shown in Fig. 6, 7, 8, 9.

(c) Rotate the coordinate through the angle $\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\rho\sqrt{s_{11}s_{22}}}{s_{11}-s_{22}}\right)$. Plot locations and the histograms of the rotated random sample as in (a).

- Solution: See Fig. 10, 11, 12, 13, 14.

(d) Calculate the sample correlation coefficient of the rotated random sample. Are the variables of the rotated population statistically independent?

- Solution: The new sample correlation coefficient is 1.1032×10^{-15}.

  From the histogram we can get the estimate probability density function (pdf) $\hat{f}(x_1, x_1), \hat{f}(x_1), \hat{f}(x_2)$ of the population respectively. For the tolerance interval $\delta = 0.1$, we calculate $\#(|f(x_1)f(x_2) \neq f(x_1, x_2)| \geq \delta) = 3.9212 \times 10^{-4}$. For the tolerance interval $\delta = 0.01$, we calculate $\#(|f(x_1)f(x_2) \neq f(x_1, x_2)| \geq \delta) = 0.0753$. With some tolerance of measurement error and noise, the statistics show that the two variables of the population are statistically independent.

Note: The rotation angle is calculated as below. Consider random variables Y_1 and Y_2 related to arbitrary random variables X_1 and X_2 by the coordinate rotation

$$\begin{aligned} Y_1 &= X_1\cos(\theta) + X_2\sin(\theta) \\ Y_2 &= -X_1\sin(\theta) + X_2\cos(\theta). \end{aligned}$$

The covariance of Y_1 and Y_2 is

$$\begin{aligned} s_{Y_1Y_2} &= \mathbb{E}\Big[(Y_1 - \bar{Y}_1)(Y_2 - \bar{Y}_2)\Big] \\ &= \mathbb{E}\Big[\{(X_1 - \bar{X}_1)\cos(\theta) + (X_2 - \bar{X}_2)\sin(\theta)\} \\ &\cdot \quad \{-(X_1 - \bar{X}_1)\sin(\theta) + (X_2 - \bar{X}_2)\cos(\theta)\}\Big] \\ &= (s_{X_2} - s_{X_1})\sin(\theta)\cos(\theta) + s_{X_1X_2}\Big[\cos^2(\theta) - \sin^2(\theta)\Big] \\ &= (s_{X_2} - s_{X_1})\sin(2\theta)/2 + s_{X_1X_2}\cos(2\theta). \end{aligned}$$

Here $s_{X_1X_2} = r_{X_1X_2}\sqrt{s_{X_1}s_{X_2}}$. If we require Y_1 and Y_2 to be uncorrelated, we must have $s_{Y_1Y_2} = 0$. Thus the coordinate rotation through the angle $\theta = \frac{1}{2}\tan^{-1}\left(\frac{2\rho\sqrt{s_{X_1X_1}s_{X_2X_2}}}{s_{X_1X_1}-s_{X_2X_2}}\right)$ is sufficient to convert correlated random variables into two uncorrelated random variables (statistically independent for Gaussian random variables).

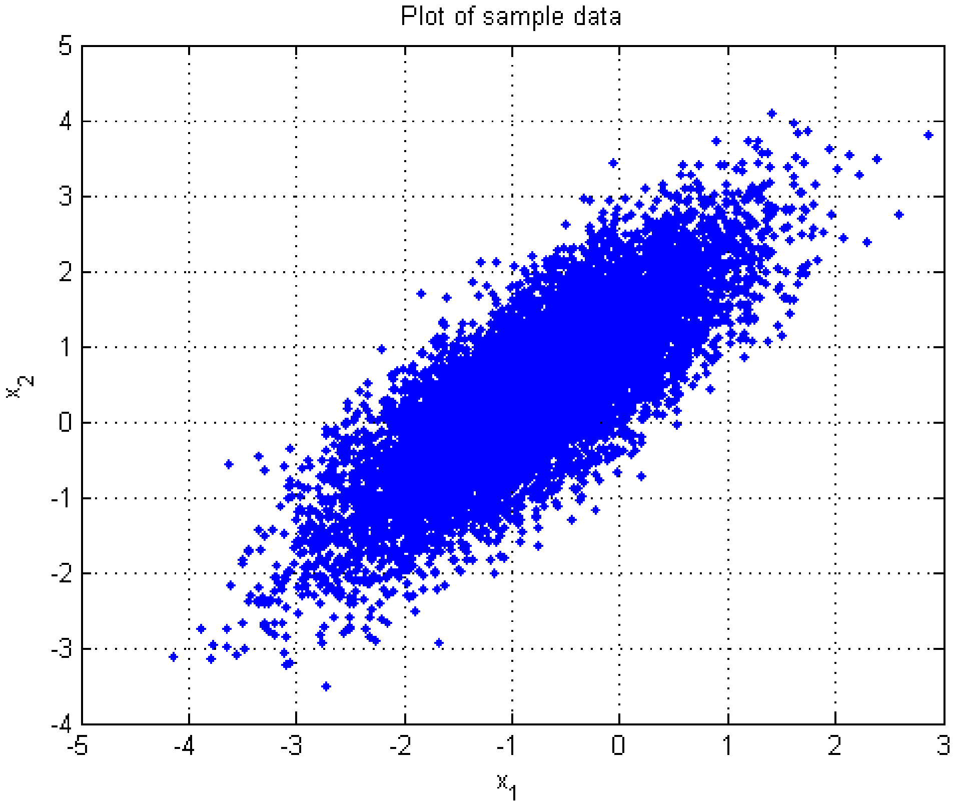

Figure 1: Distribution of The Random Sample in $\mathbb{R}^2$

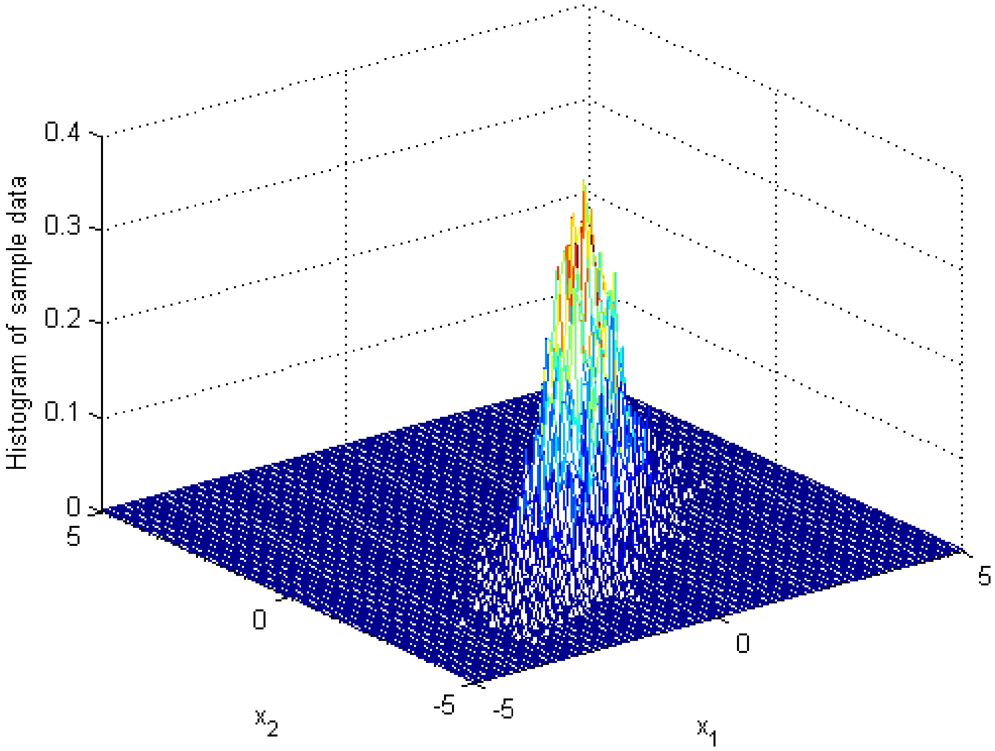

Figure 2: Histogram of The Random Sample

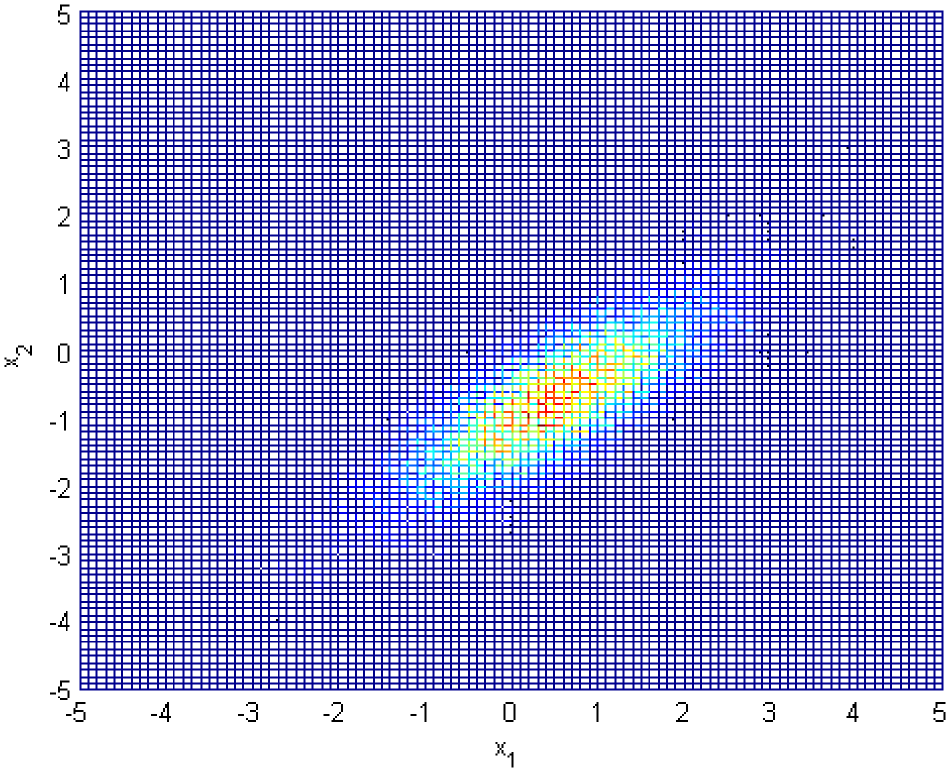

Figure 3: Contour of The Random Sample

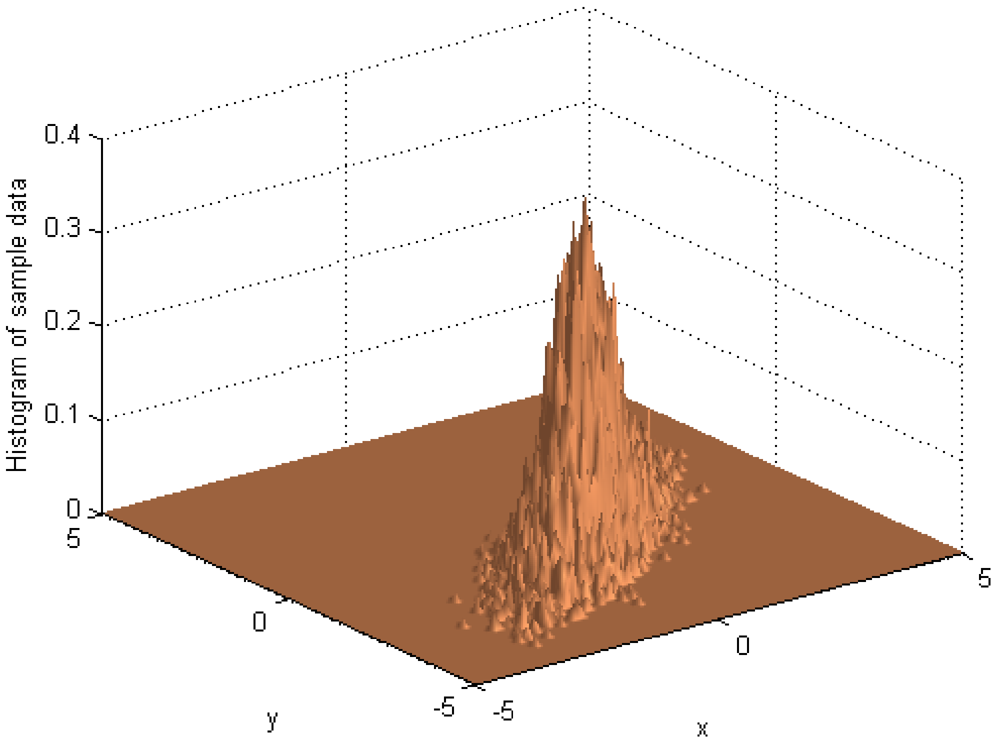

Figure 4: Histogram of The Random Sample

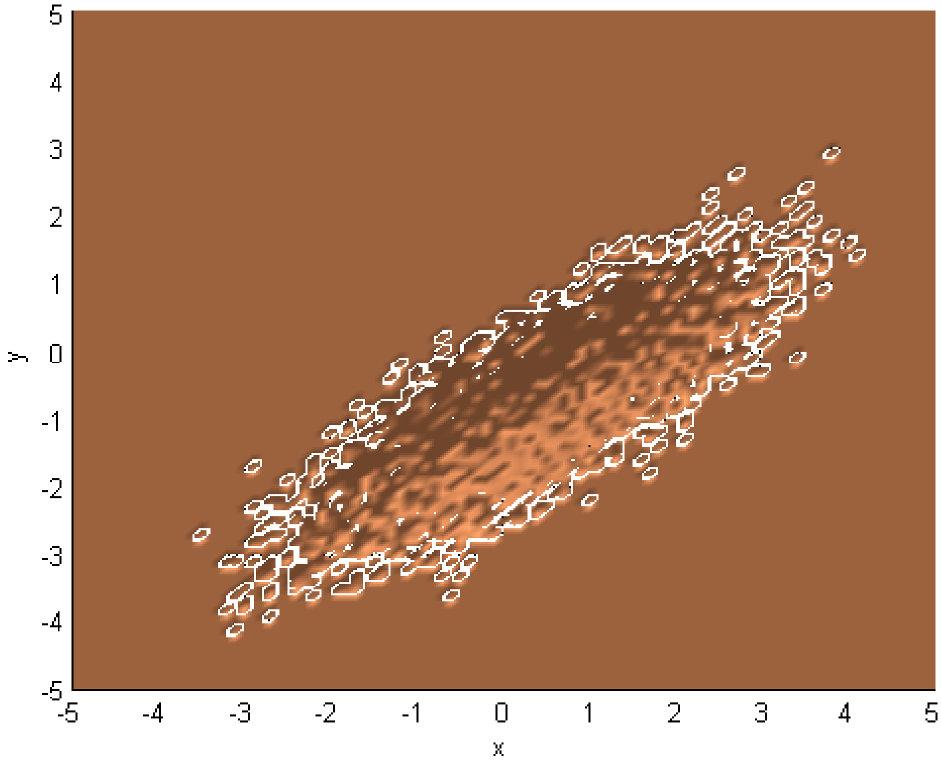

Figure 5: Contour of The Random Sample

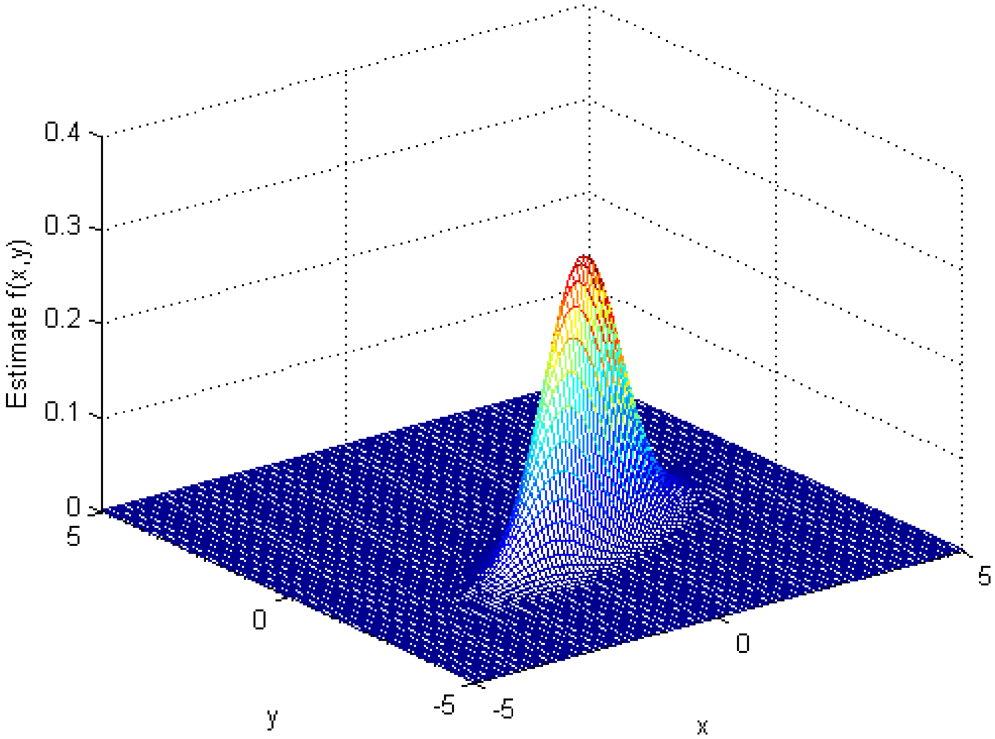

Figure 6: Histogram of The Population

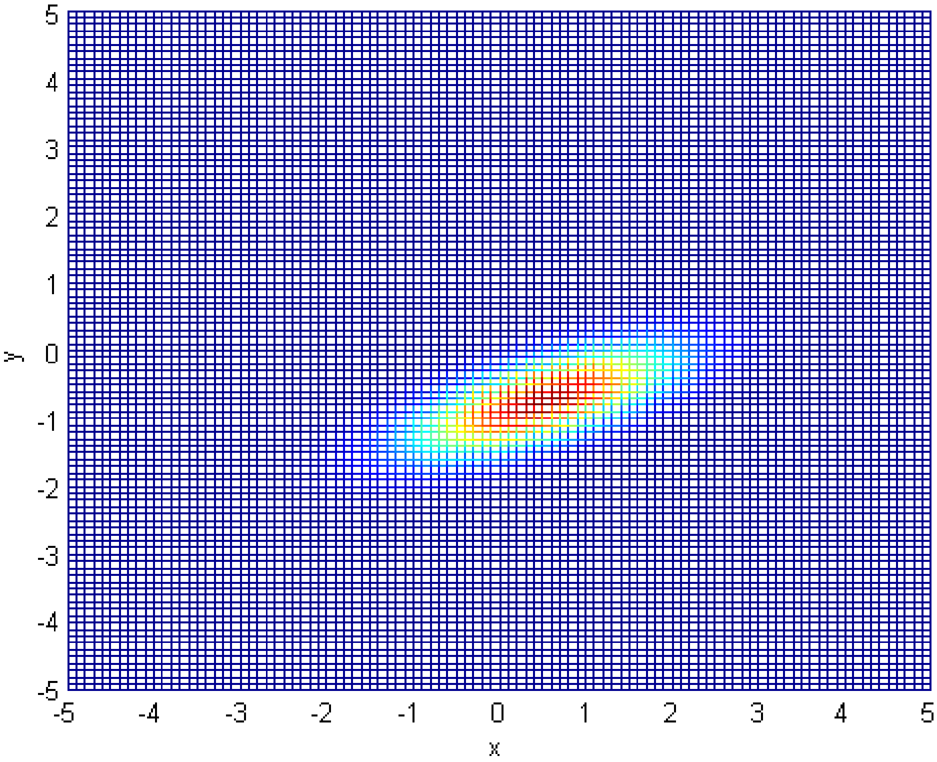

Figure 7: Contour of The Population

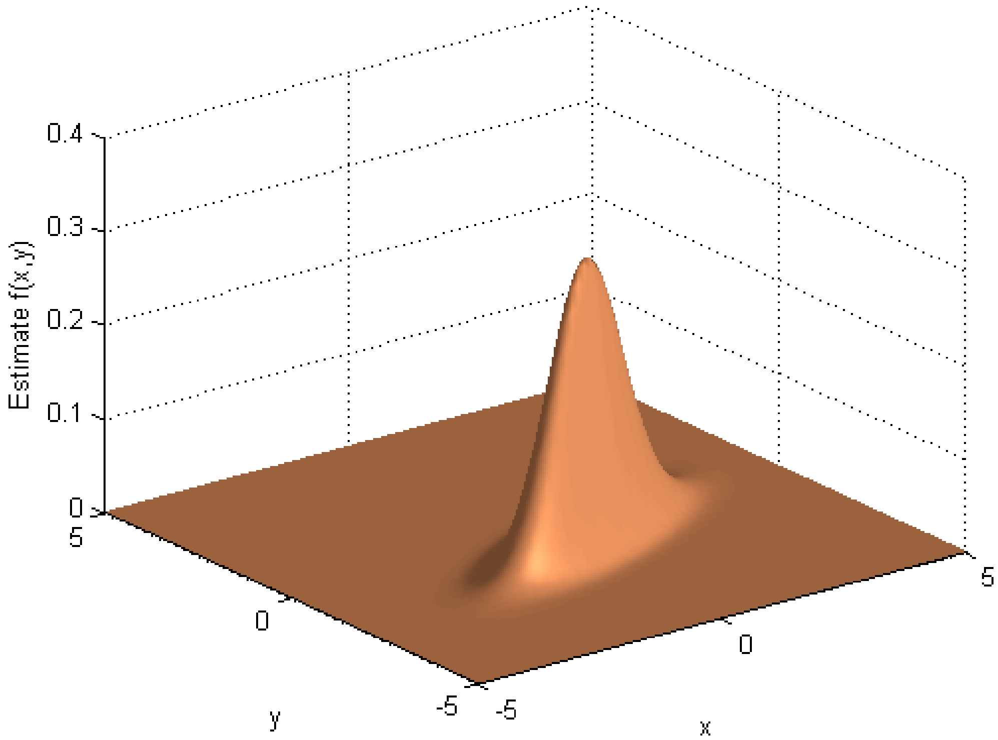

Figure 8: Histogram of The Population

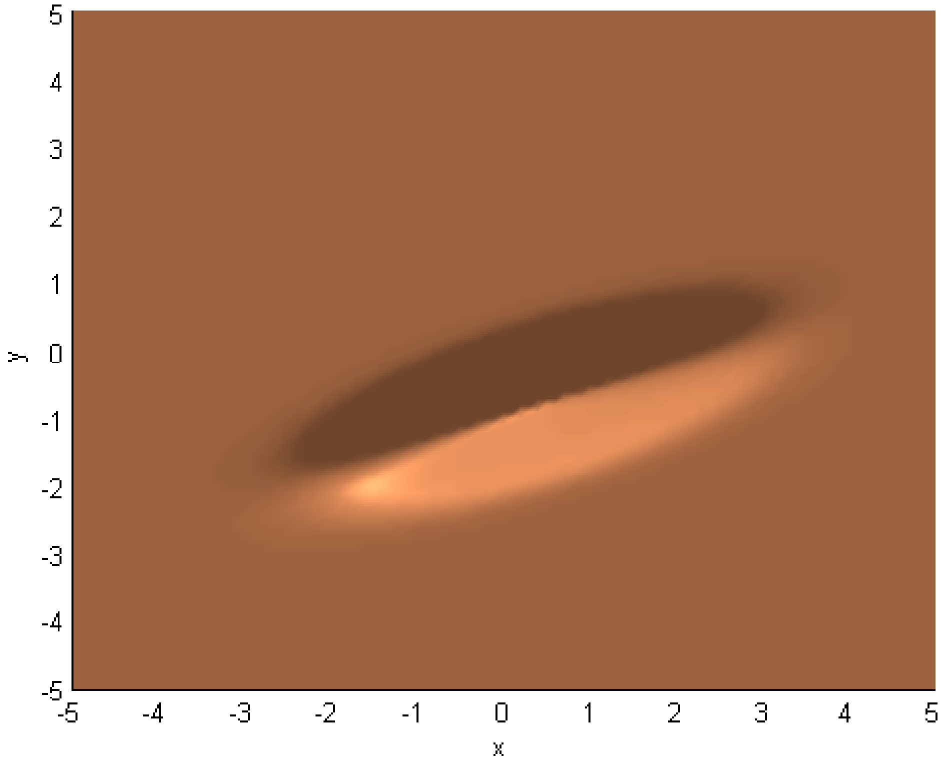

Figure 9: Contour of The Population

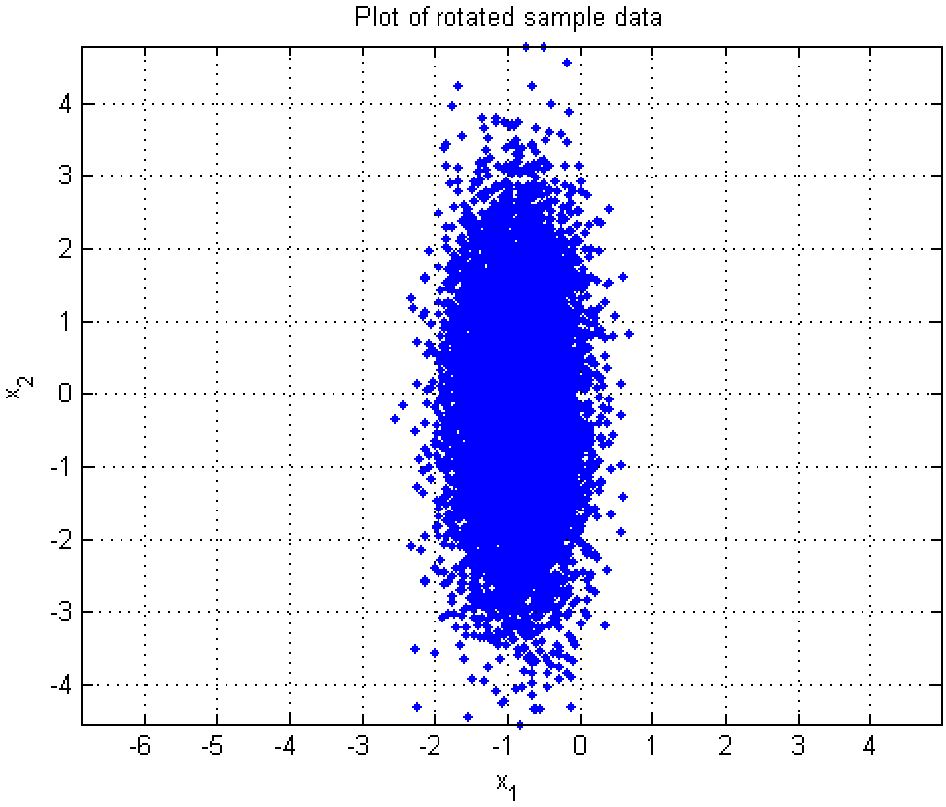

Figure 10: Distribution of The Rotated Random Sample in $\mathbb{R}^2$

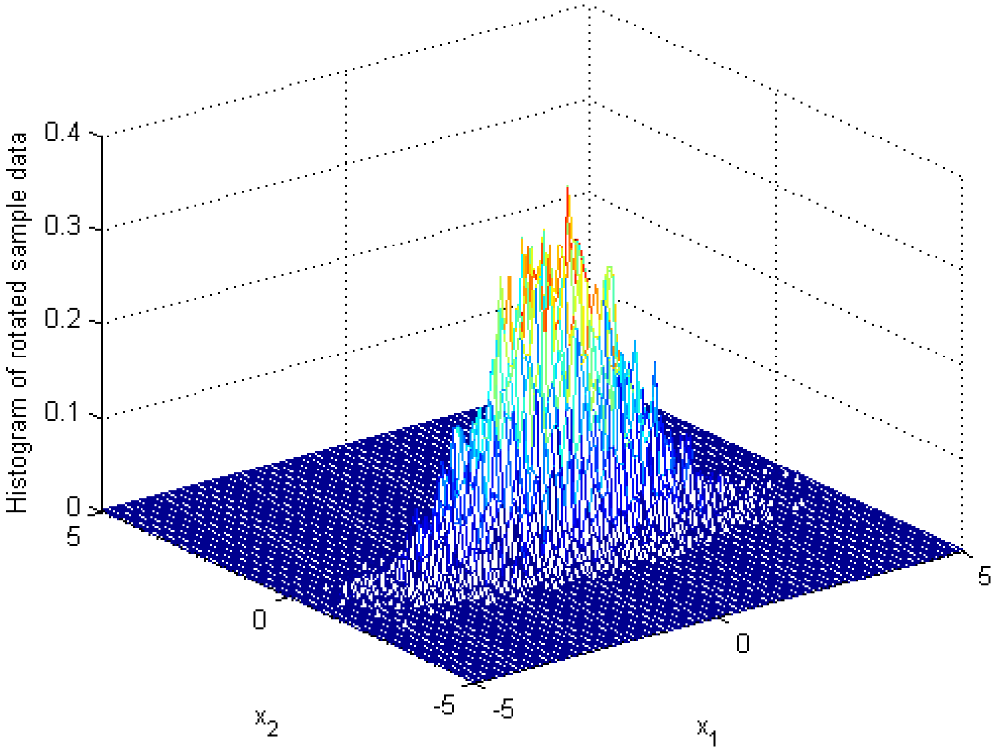

Figure 11: Histogram of The Rotated Random Sample

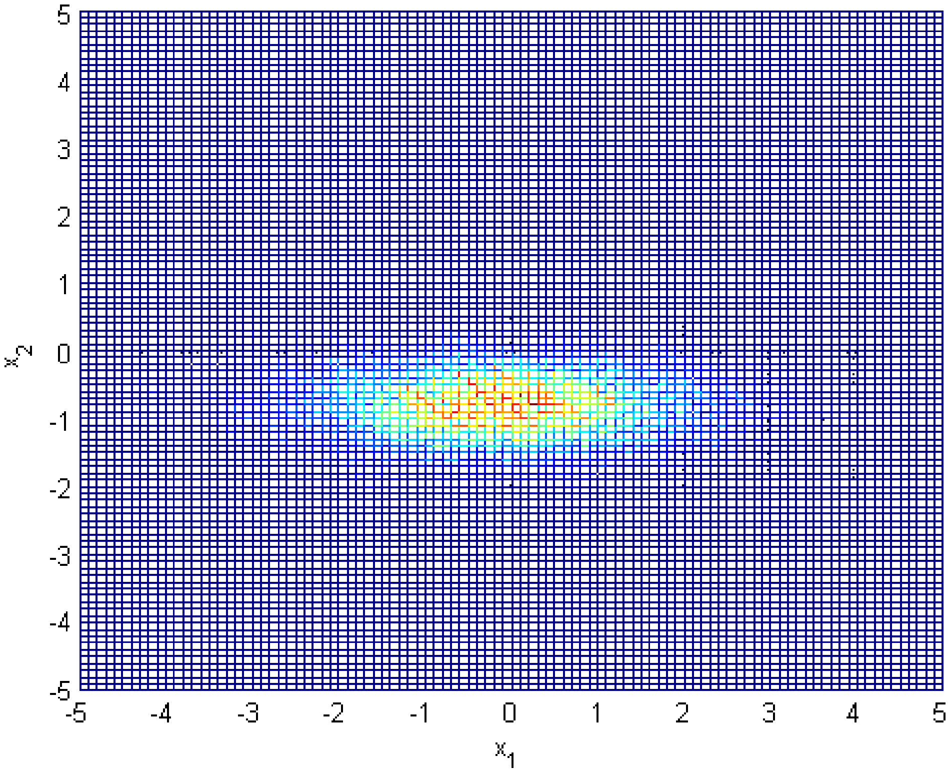

Figure 12: Contour of The Rotated Random Sample

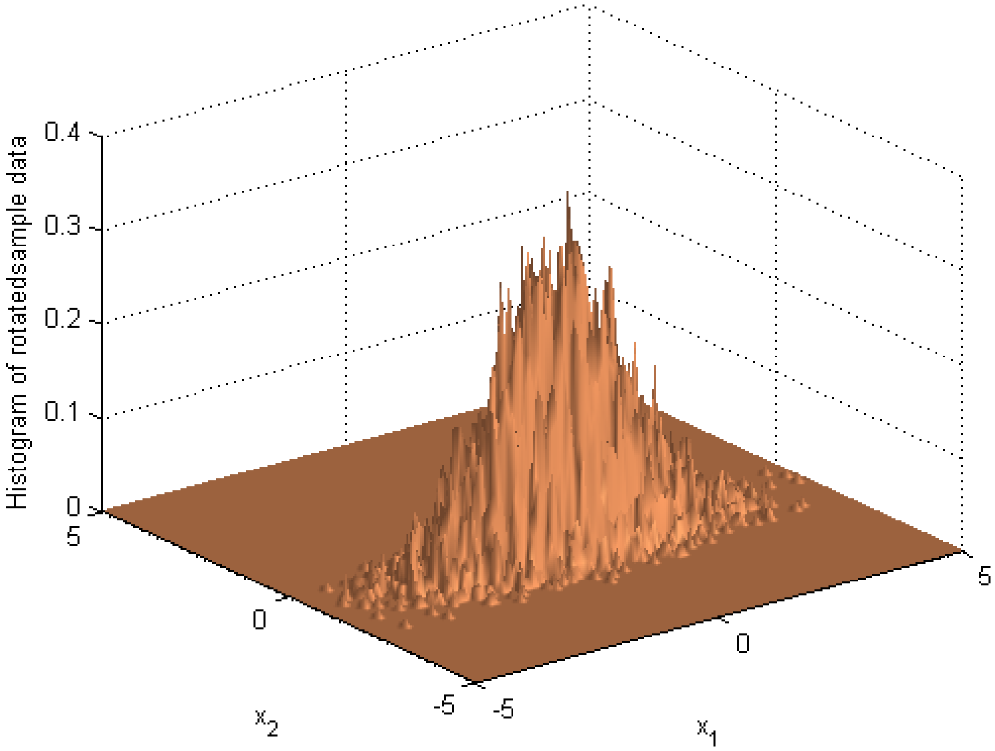

Figure 13: Histogram of The Rotated Random Sample

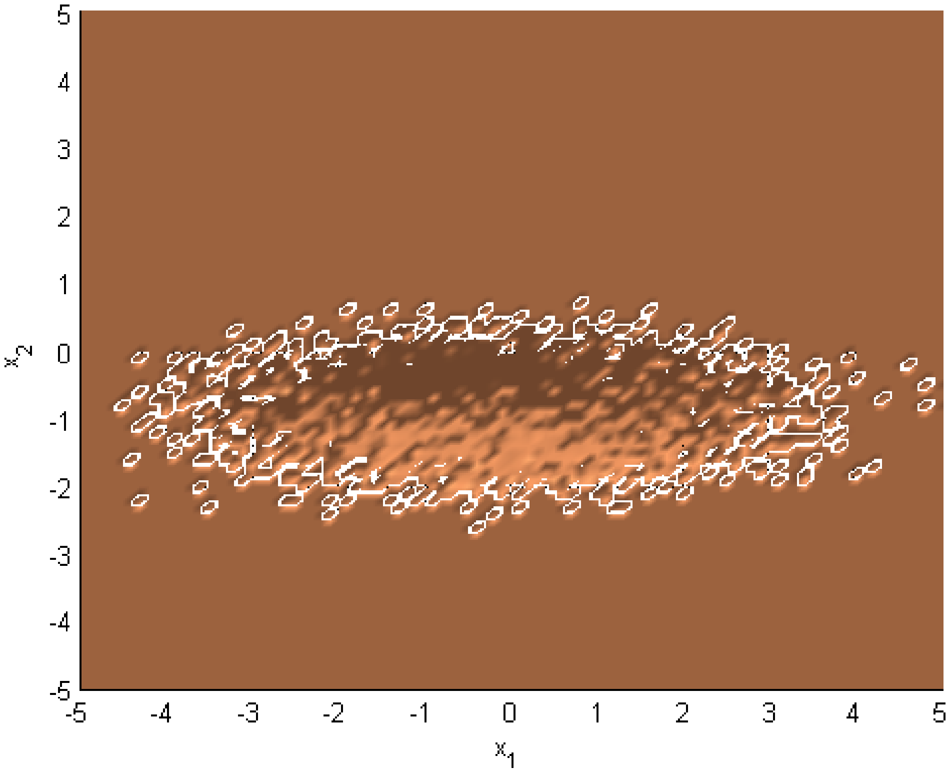

Figure 14: Contour of The Rotated Random Sample

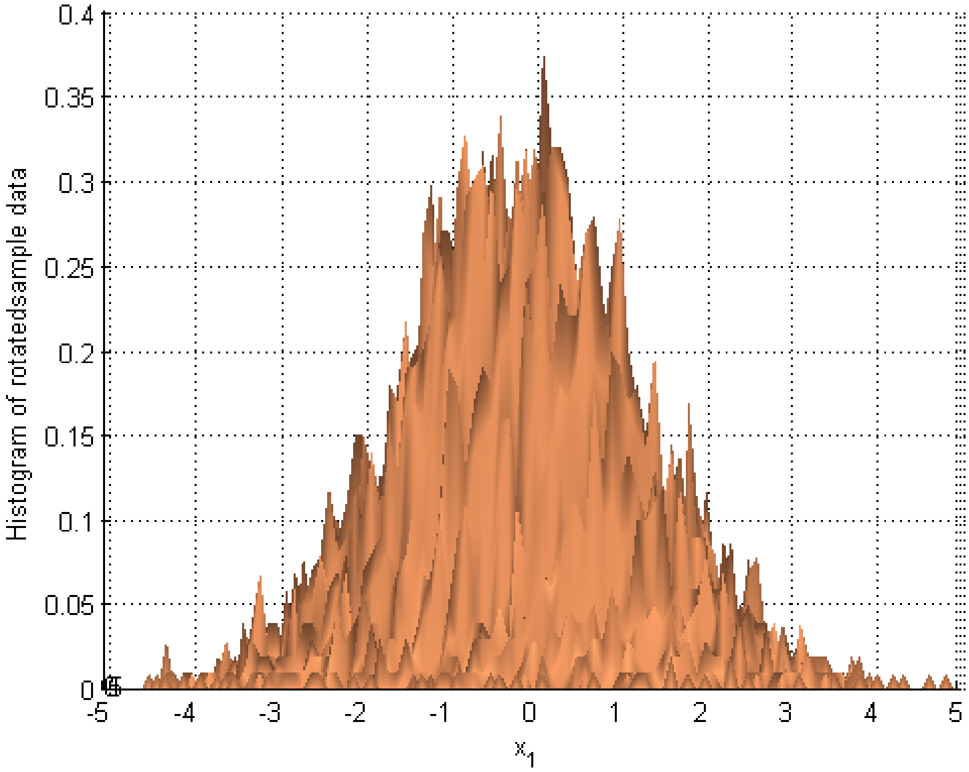

Figure 15: Histogram of x_1

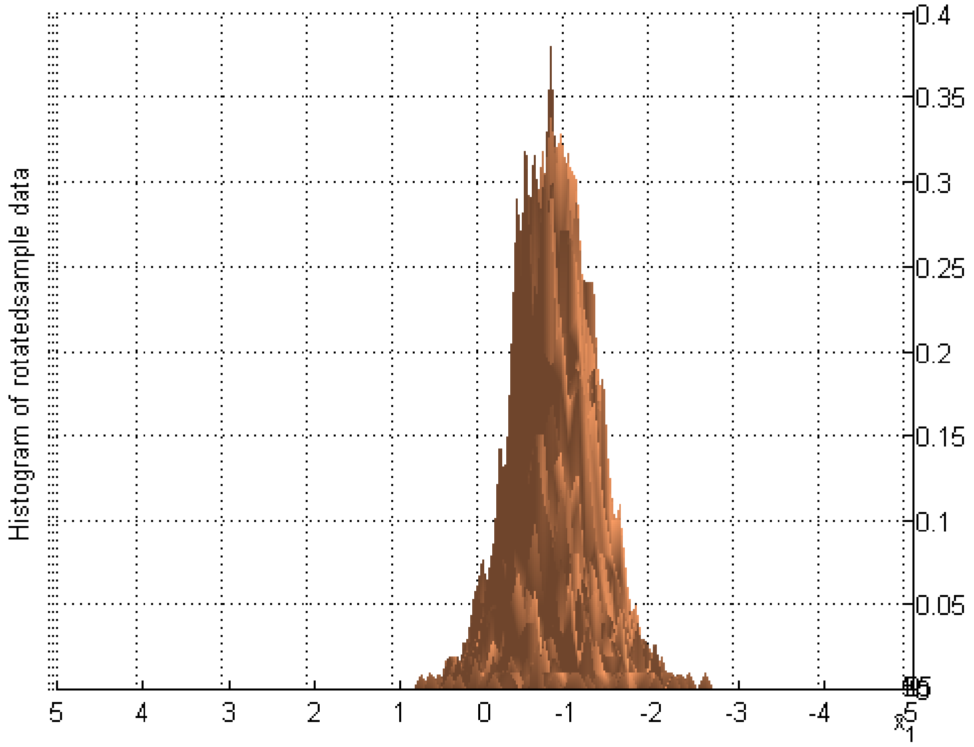

Figure 16: Histogram of x_2